DOI:10.11931/guihaia.gxzw201810027

巨桉叶绿体基因组密码子偏好性分析

王鹏良[1,2]，吴双成[2]，杨利平[3]，王华宇[2]，陈乃明[3]，张照远[1,*]

（1. 广西优良用材林资源培育重点实验室，广西壮族自治区林业科学研究院，南宁 530002；2. 广西北部湾海洋生物多样性养护重点实验室，钦州学院，广西 钦州 535011；3. 钦州市植物生物技术重点实验室，广西钦州市林业科学研究所，广西 钦州 535099）

摘要：为了提高基因表达效率从而利用叶绿体基因工程改良巨桉重要性状，该文以巨桉叶绿体基因组序列为材料，选取其中长于 300nt 且以 AUG 为起始密码子的 43 个非重复基因为研究对象，采用 CodonW1.4.2 软件分析巨桉叶绿体基因组的密码子使用偏好性。分析结果表明：第 3 位密码子的平均 GC 含量为 27.97%；ENC 的变化范围为 39.49~61.00，平均为 47.04；RSCU>1 的密码子有 31 个，其中 29 个以 A/U 结尾；中性分析显示 GC12 与 GC3 无显著相关，回归分析也未达到显著性水平；ENC-plot 分析发现大部分基因落在曲线上或附近；对应分析表明第一轴的贡献率为 17.68%，第二轴的贡献率为 11.49%，第三、四轴的贡献率分别为 8.00%和 5.76%，前四轴累计贡献率达 42.93%，第一轴与 GC、ENC、CAI 达到极显著相关；上述分析结果表明，巨桉叶绿体基因组的密码子偏好较弱，密码子第 3 位偏好以 A 或 U 结尾；选择和突变在巨桉叶绿体基因组密码子偏好中起相对均衡的作用；最终确定 UUG、CUU、GUU、UCC、UCA、ACA、UAU、UAA、CAU、AAU、AGA 和 GGA 12 个高频高表达密码子为最优密码子。该研究为转化叶绿体基因密码子优化从而提高表达效率改良巨桉目标性状奠定了坚实基础。

关键词：巨桉，叶绿体，基因组，密码子偏好性

Analysis of codon bias of chloroplast genome in *Eucalyptus grandis*

WANG Pengliang[1,2], WU Shuangcheng[2], YANG Liping[3], WANG Huayu[2], CHEN Naiming[3], ZHANG Zhaoyuan[1*]

(1. *Guangxi Forestry Research institute, Guangxi Key Laboratory of Superior Timber Trees Resource Cultivation*, Nanning 530002, China; 2. *Qinzhou University, Guangxi Key Laboratory of Beibu Gulf Marine Biodiversity Conservation*, Qinzhou 535011, Guangxi, China; 3. *Qinzhou Forestry Research institute, Qinzhou Key Laboratory of Plant biotechnology*, Qinzhou 535099, Guangxi, China)

Abstract: In order to increase the expression efficiency of genes in chloroplast for future

基金项目：广西科技重大专项(桂科 AA17204087-3)；广西主要用材林资源高效培育与利用人才小高地专项(桂财社函[2018]112 号)；广西优良用材林资源培育重点实验室自主项目(14-A-03-01)；广西优良用材林资源培育重点实验室开发课题基金(12A0301)；钦州市科学研究与技术开发项目(20137003) [Supported by Major Scientific and Technological Program j in Guangxi (AA17204087-3), the Department of Human Resources and Social Security of Guangxi Zhuang Autonomous Region ([2018]112); Autonomous Program of Guangxi Key Laboratory of Superior Timber Trees Resource Cultivation (14-A-03-01); Open project of Guangxi Key Laboratory of Superior Timber Trees Resource Cultivation (12A0301); Program of Science and Technology in Qinzhou (20137003)]。

作者简介：王鹏良（1978 -），男，浙江新昌人，博士，高级工程师，主要从事植物遗传育种研究，(E-mail) pengliang_wang@163.com; pengliang_wang@qzhu.edu.cn。

***通讯作者**：张照远，博士，高级工程师，研究方向为林木遗传育种，(E-mail) zzy3564@163.com。

improvement of important traits in *Eucalyptus grandis*, analysis of codon bias was carried out using CodonW 1.4.2 software, with chloroplast genome of *Eucalyptus grandis* as a material and 43 non-repeated genes beginning with AUG as objects. The results indicated that the average of GC content in 3rd position was 27.97%; ENC ranged from 39.49 to 61.00 with an average of 47.04; there were 31 codons whose RSCUs were more than 1.00 in the chloroplast genome; of them, 29 codons ended with A/U; neutral plot analysis showed correction and regression analysis between GC12 and GC3 were not significant; ENC-plot revealed most genes located along or near the standard curve; correspondence analysis indicated the 1st axis accounted for 17.68% contribution, the 2nd axis 11.49%, the rest axes accounted for 8.00% and 5.76% and the first four axes accounted for 42.93% in total; the correction between the 1st axis and GC3S was not significant, however the negative correction between 1st axis and ENC was significant. The results mentioned above revealed that the codon bias level was low in the chloroplast genome and the codon always end with A/U and codon bias might be determined by both mutation and selection nearly equally. Finally, the codons that were not only highly expressed but frequently were determined as the optimal codons including UUG, CUU, GUU, UCC, UCA, ACA, UAU, UAA, CAU, AAU, AGA and GGA. This study will provide a solid foundation for codon optimization of the genes transformed into chloroplast genome and future increasing the expression efficiency for improvement of important traits.

Keywords: *Eucalyptus grandis,* chloroplast, genome, codon bias

巨桉(*Eucalyptus grandis*)原产于澳大利亚，为桃金娘科桉属中的一个多年生木本树种。因其生长迅速，树形通直，树体高大，巨桉被引种至世界各地广泛种植，成为各国重要的外来树种(祁述雄，2006; 陈少雄等，2018)。因此，研究人员在引种驯化的基础上开展了种源/家系/单株不同性状变异研究；结果表明，巨桉的抗寒性不足(刘建等, 2009)，易受瘿姬小蜂感染(张照远等，2016)，不同遗传资源在生长和形质方面也存在较大差异(吴世军等, 2016; 张捷等, 2016)。

基因工程技术育种与传统育种技术相比具有针对性强，周期短，效率高等明显优势(王关林和方宏筠, 2014)。叶绿体基因工程具有明显的高效表达，并能有效控制转化基因的扩散等特点，是极为理想的转化方式(Daniell & Chase, 2004)。密码子被称为第二套遗传密码(Nelson & Cox, 2017; Hanson & Coller, 2018)；密码子使用的选择不仅影响基因的表达(Zhou et al, 2016)，也影响基因相应的功能(Hershberg & Petrov, 2008)。不同物种间叶绿体基因组的密码子偏好存在较大差异(Zhou et al, 2008; 刘慧等, 2017; 王鹏良等, 2018; 王文斌等, 2018)。本文旨在分析巨桉叶绿体基因组密码子偏好性的特征，并确定其最优密码子，为巨桉叶绿体基因工程的开展和遗传改良奠定基础。

1. 材料和方法

1.1 序列

从 NCBI 网站的细胞器基因组网页中搜索巨桉的拉丁名：*Eucalyptus grandis* 找到巨桉的叶绿体基因组（https://www.ncbi.nlm.nih.gov/nuccore/NC_014570.1），下载其 Fasta 格式的全基因组和基因编码序列(Coding sequences)。巨桉叶绿体基因组总长为 160137bp，共含有 75 个基因。为了降低误差，本文选用其中以 AUG 为起始密码子且长度超过 300nt 的 43 条非重复序列用于密码子偏好性分析。

1.2 数据分析

1.2.1 密码子偏好参数计算

以所选的 43 个非重复基因的编码序列为对象，采用 CodonW1.4.2 软件分析密码子偏好参数：同义密码子相对使用度（RSCU, relative synonymous codon usage）、有效密码子数目（ENC, effective number of codon），其最小理论值为 20，说明每个氨基酸都只有一个密码子，最大理论值为 61，说明所有密码子都均等使用、密码子适应指数(CAI, codon adaption index)，变化范围为 0 ~ 1，值越大偏性越强、密码子偏好性指数(CBI, codon bias index)，最优密码子使用频率(FOP, frequency of optimal codons)，该基因表达为蛋白质的疏水性(Gravy)及不同位置的 GC 含量，包括 GC1，GC2，GC3，GC3S，GC12 和 GC，分别代表密码子中第 1，2，3 位的 GC 含量，第 3 位同义密码子 GC 含量，第 1，2 位密码子平均 GC 含量和密码子总体的 GC 含量。

1.2.2 中性绘图分析

为了初步确定影响密码子偏好的因素，中性绘图分析根据 GC1 和 GC2 的信息计算两者的平均值 GC12 作为纵坐标，以 GC3 为横坐标，以散点图的形式在坐标中定位各基因的位置，根据基因的坐标信息与坐标对角线的关系，若基因位于对角线上，则表明基因受突变作用；若基因不位于对角线，则表明该基因收到选择的影响，从而判断造成密码子的使用偏好的因素。

1.2.3 ENC-plot 绘图

为了进一步确定影响密码子偏好的因素，ENC-plot 绘图以 ENC 为纵坐标，以 GC3S 为横坐标建立坐标系，将各基因定位在该坐标中形成散点图。再在坐标系中添加 ENC 的标准曲线，标准曲线方程(Wright, 1990)为：

$$ENC_{exp} = 2 + GC3S + \frac{29}{GC3S^2 + (1-GC3S)^2} \quad (1)式$$

ENC 比值的公式为：

$$ENC_{Ratio} = \frac{ENC_{exp} - ENC_{obs}}{ENC_{exp}} \quad (2)式$$

根据散点图和 ENC 比值的分布结果；若偏离标准曲线，则表明受到选择作用；若在标准曲线上，则只是受到突变作用从而推断造成密码子偏好的可能原因。

1.2.4 对应分析

对应分析是一种对原始数据采用适当的标度方法，将变量和样本分析结合起来，同时得到两方面的结果，在同一因子平面上对变量和样本一起进行分类，从而揭示样本和变量间的内在联系。利用 CodonW 软件将对应分析用于巨桉叶绿体基因组密码子分析，从而揭示巨桉叶绿体基因组密码子使用的规律。

1.2.5 最优密码子的确定

为了确定最优密码子，本文以 ENC 参数为标准对所有参试基因按从大到小的顺序排列，分别从 ENC 最高和最低两端都选取所有参试基因的 10%，建立高表达和低表达库。

$$\Delta RSCU = RSCU_{highexpressiongenes} - RSCU_{lowexpressiongenes} \quad (3)式$$

将高表达库与低表达库的同义密码子相对使用度的差值(ΔRSCU)高于 0.08 且同义密码子相对使用度(RSCU)高于 1 的密码子确定为最优密码子(李娟和薛庆中, 2005; 续晨等, 2010; 杨国锋等, 2015; 王鹏良等, 2018)。

2.结果与分析

2.1 密码子组成分析

为了更加准确分析密码子偏好性，本研究选取了巨桉叶绿体基因组中以 AUG 为起始密码子且编码区序列长度超过 300 nt 的 43 个非重复基因的编码序列为研究对象，采用 codonW 软件对参试基因开展密码子相关参数的计算和分析。结果（表 1）表明：不同基因密码子不同位置的 GC 含量并不相同，第 1、2、3 位密码子的 GC 含量的变化范围分别为 34.20%~58.90%，27.90%~58.70%，20.20%~37.00%，其平均值分别为 47.40%，39.47%，27.97%；第 1，2 位的 GC 含量明显高于第 3 位。ENC 的范围在 39.49~61.00 之间，平均值为 47.04。CAI 的范围为 0.082~0.301，平均值为 0.1714。CBI 的范围为-0.222~0.196，平均值为-0.092。FOP 的范围为 0.263~0.532，其平均值为 0.356。蛋白质的 Gravy 变化范围为-0.704~1.102，平均值为 0.017。

表 1 巨桉叶绿体基因组密码子主要参数

Table1 Main parameters in chloroplast genomics of *Eucalyptus grandis*

Gene	GC1	GC2	GC3	GC	GC3S	ENC	CAI	CBI	FOP	Gravy
*psb*A	0.504	0.433	0.328	0.422	0.284	41.24	0.301	0.196	0.532	0.341
*mat*K	0.402	0.312	0.276	0.330	0.257	51.29	0.167	-0.141	0.331	-0.172
*atp*A	0.558	0.397	0.248	0.401	0.232	44.61	0.203	-0.037	0.393	-0.055
*atp*F	0.478	0.332	0.348	0.386	0.330	48.37	0.155	-0.106	0.358	-0.345
*atp*I	0.482	0.369	0.263	0.371	0.235	44.97	0.168	-0.075	0.353	0.625
*rps*2	0.445	0.432	0.271	0.383	0.236	47.19	0.167	-0.174	0.316	-0.282
*rpo*B	0.503	0.377	0.280	0.387	0.259	48.77	0.149	-0.121	0.341	-0.286
*psb*D	0.527	0.434	0.317	0.426	0.272	44.02	0.263	0.083	0.465	0.359
*psb*C	0.537	0.461	0.302	0.433	0.263	45.82	0.196	-0.003	0.406	0.265
*psa*B	0.487	0.429	0.308	0.408	0.263	47.83	0.182	-0.108	0.356	0.117
*psa*A	0.527	0.433	0.324	0.428	0.283	50.26	0.197	-0.099	0.358	0.259
*ycf*3	0.477	0.387	0.333	0.399	0.304	61.00	0.151	-0.178	0.335	-0.502
*rps*4	0.503	0.383	0.263	0.383	0.245	51.59	0.155	-0.053	0.372	-0.606
*ndh*J	0.525	0.387	0.285	0.399	0.242	47.52	0.162	-0.180	0.302	-0.273
*ndh*C	0.475	0.333	0.259	0.356	0.198	46.03	0.182	-0.104	0.324	1.094
*atp*E	0.518	0.406	0.255	0.393	0.227	50.16	0.164	-0.054	0.375	-0.068
*atp*B	0.565	0.416	0.284	0.421	0.261	45.37	0.202	-0.006	0.406	-0.035
*rbc*L	0.575	0.435	0.288	0.433	0.260	46.54	0.265	0.071	0.464	-0.284
*ycf*4	0.462	0.403	0.342	0.402	0.309	50.13	0.183	-0.032	0.389	0.240
*cem*A	0.394	0.279	0.297	0.323	0.251	45.87	0.188	-0.046	0.377	0.248
*pet*A	0.531	0.363	0.315	0.403	0.303	52.23	0.189	-0.048	0.382	-0.114
*rps*18	0.376	0.416	0.248	0.347	0.232	43.41	0.106	-0.153	0.313	-0.614
*rpl*20	0.351	0.453	0.231	0.345	0.211	45.47	0.082	-0.222	0.263	-0.551
*clp*P	0.589	0.364	0.323	0.426	0.283	53.46	0.181	-0.123	0.332	0.088
*psb*B	0.551	0.460	0.282	0.431	0.241	45.75	0.185	-0.073	0.372	0.113
*pet*B	0.493	0.414	0.274	0.394	0.216	39.49	0.216	-0.046	0.372	0.582
*rpo*A	0.454	0.326	0.267	0.349	0.247	48.93	0.157	-0.122	0.345	-0.334
*rps*11	0.522	0.587	0.202	0.437	0.173	42.79	0.152	-0.140	0.331	-0.415
*rps*8	0.388	0.403	0.328	0.373	0.302	47.18	0.108	-0.042	0.380	-0.325
*rpl*14	0.533	0.386	0.246	0.388	0.227	44.66	0.165	-0.070	0.361	-0.043
*rpl*16	0.503	0.525	0.222	0.417	0.167	41.74	0.121	-0.100	0.357	-0.460

*rps*3	0.476	0.347	0.259	0.361	0.227	43.47	0.154	-0.150	0.329	-0.318
*rpl*22	0.419	0.369	0.269	0.352	0.220	47.47	0.198	-0.075	0.393	-0.704
*ycf*2	0.414	0.344	0.370	0.376	0.343	53.16	0.158	-0.140	0.337	-0.435
*ndh*B	0.421	0.384	0.317	0.375	0.280	47.80	0.164	-0.084	0.354	0.685
*rps*7	0.529	0.458	0.246	0.411	0.215	43.44	0.188	-0.055	0.389	-0.593
*ndh*F	0.370	0.364	0.221	0.319	0.181	43.50	0.133	-0.210	0.282	0.535
*ccs*A	0.342	0.379	0.273	0.331	0.221	47.49	0.139	-0.198	0.295	0.551
*ndh*D	0.406	0.366	0.270	0.347	0.230	46.76	0.135	-0.129	0.316	0.788
*ndh*E	0.396	0.337	0.248	0.327	0.224	49.43	0.142	-0.198	0.286	0.734
*ndh*G	0.438	0.341	0.261	0.347	0.231	45.73	0.140	-0.181	0.278	1.102
*ndh*I	0.426	0.378	0.228	0.343	0.194	42.59	0.203	-0.112	0.350	-0.036
*ndh*H	0.509	0.369	0.254	0.377	0.202	48.27	0.155	-0.113	0.338	-0.145
平均	0.474	0.395	0.280	0.383	0.246	47.04	0.171	-0.092	0.356	0.017

密码子参数的相关分析（表 2）表明，GC1 与 GC2 为显著相关，其相关系数为 0.363；GC1 与 GC3 相关不显著，GC2 与 GC3 相关也不显著；同时 GC 含量与 GC1 和 GC2 极显著相关，与 GC3 无显著相关。ENC 与 GC1 不相关，与 GC2 显著负相关，与 GC3 极显著正相关，其相关系数为 0.521。GC1 和 GC 两个参数与 CAI, CBI 和 FOP 极显著相关，GC3 与 CAI、CBI 和 FOP 呈显著相关；Gravy 与其余的密码子参数均无显著相关；密码子数目（N）与 GC3 极显著相关外，不与 ENc 和 CAI 等其他参数显著相关。

表 2 密码子参数的相关性分析

Table 2 Correlation analysis of the parameters of coding sequences' codon usage

Item	GC1	GC2	GC3	GC	GC3S	ENC	CAI	CBI	FOP	Gravy
GC2	0.363*									
GC3	0.117	-0.258								
GC	0.857**	0.669**	0.300							
GC3S	0.114	-0.282	0.950**	0.266						
ENc	0.008	-0.365*	0.521**	0.000	0.584**					
CAI	0.551**	0.084	0.320*	0.502**	0.225	-0.184				
CBI	0.513**	0.213	0.340*	0.555**	0.295	-0.263	0.813**			
Fop	0.526**	0.241	0.380*	0.593**	0.338*	-0.188	0.834**	0.997**		
Gravy	-0.120	-0.270	0.040	-0.204	-0.091	-0.188	0.177	0.045	-0.091	
N	-0.030	-0.122	0.399**	0.062	0.397	0.234	0.071	-0.018	0.023	-0.025

注：*表示在 0.05 水平上显著相关；**表示在 0.01 水平上显著相关。

Note: * means significant correlation at 0.05 level; ** means significant correlation at 0.01 level.

RSCU 分析（表 3）表明，RSCU 大于 1.00 的密码子数目为 31 个。其中，以 U 结尾的密码子有 16 个，以 A 结尾的密码子有 13 个，以 G 和 C 结尾的密码子分别为 1 个。以 A 或 U 结尾的密码子占全部的 93.54%。

表 3 巨桉叶绿体基因同义密码子相对使用度

Table3 Relative synonymous codon usage analysis of genes on chloroplast genome in *Eucalyptus grandis*

氨基酸	密码子	数目	RSCU	氨基酸	密码子	数目	RSCU

Amino acid	Codon	Number		Amino acid	Codon	Number	
Phe	UUU	584	1.28	Ser	UCU	340	1.8
	UUC	329	0.72		UCC	192	1.02
Leu	UUA	562	1.99		UCA	202	1.07
	UUG	321	1.14		UCG	110	0.58
	CUU	366	1.3	Pro	CCU	268	1.68
	CUC	105	0.37		CCC	107	0.67
	CUA	241	0.85		CCA	185	1.16
	CUG	99	0.35		CCG	77	0.48
Ile	AUU	657	1.48	Thr	ACU	337	1.68
	AUC	266	0.6		ACC	156	0.78
	AUA	412	0.93		ACA	222	1.11
Met	AUG	385	1		ACG	86	0.43
Val	GUU	309	1.44	Ala	GCU	455	1.88
	GUC	93	0.43		GCC	143	0.59
	GUA	339	1.58		GCA	269	1.11
	GUG	119	0.55		GCG	102	0.42
Tyr	UAU	475	1.63	Cys	UGU	120	1.45
	UAC	107	0.37		UGC	46	0.55
TER	UAA	22	1.57	TER	UGA	11	0.79
	UAG	9	0.64	Trp	UGG	312	1
His	CAU	312	1.57	Arg	CGU	226	1.52
	CAC	85	0.43		CGC	58	0.39
Gln	CAA	424	1.5		CGA	214	1.44
	CAG	140	0.5		CGG	50	0.34
Asn	AAU	532	1.53	Ser	AGU	224	1.19
	AAC	164	0.47		AGC	65	0.34
Lys	AAA	521	1.51	Arg	AGA	244	1.64
	AAG	169	0.49		AGG	102	0.68
Asp	GAU	504	1.62	Gly	GGU	405	1.42
	GAC	118	0.38		GGC	119	0.42
Glu	GAA	610	1.49		GGA	450	1.57
	GAG	209	0.51		GGG	169	0.59

2.2 中性绘图分析

巨桉叶绿体基因中性绘图表明，GC12 的变化范围为 33.65%~55.45%，GC3 的变化范围为 20.20%~37.00%，GC12 与 GC3 未达到显著水平，说明 GC12 与 GC3 相关性弱。突变对密码子第 1、2 位和第 3 位碱基组成有着不同的影响。假如完全由随机突变造成的，那么基因应该在对角线上，图 1 中绝大多数基因都分布于对角线上方，GC12 均高于 GC3，绝大多数基因所在的位点高于对角线，这说明选择在密码子偏好中起主要作用。

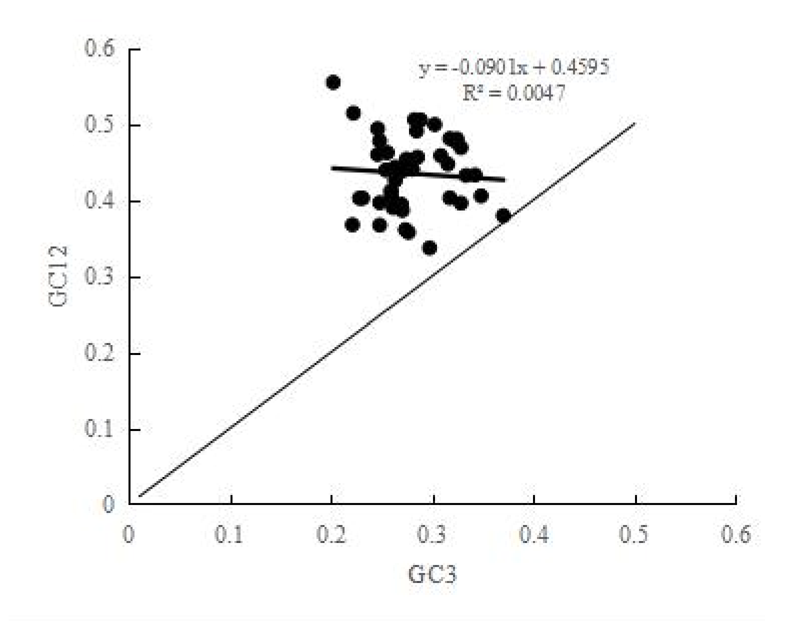


图 1 巨桉叶绿体基因中性绘图分析

Fig.1 Neutrality plot analysis of genes on chloroplast of *Eucalyptus grandis*

2.3 ENC-plot 分析

ENC-plot 绘图以 ENC 为 *y* 轴，GC3S 为 *x* 轴建立坐标系，将所有参试基因定位于该坐标系中，同时根据公式(1)添加标准曲线。ENC-plot 分析结果（图 2）表明，尽管有一小部分偏离标准曲线，大多数基因位于标准曲线附近。为了更加准确反映差异，本文根据(1)式求出 ENC 的理论值，再根据(2)式求算出 ENC 比值。在此基础上分析所有参试基因的 ENC 频数分布（表 4），统计结果表明，51.16%的基因分布在-0.05~ 0.05 之间，34.88%的基因分布在 0.05~ 0.15 之间。9.30%的基因分布在-0.15~ -0.05 之间，另有 2.33%的基因分布在-0.25~ -0.15 和 0.15~ 0.25 之间。这说明突变对巨桉叶绿体基因组密码子偏好的形成起重要作用。

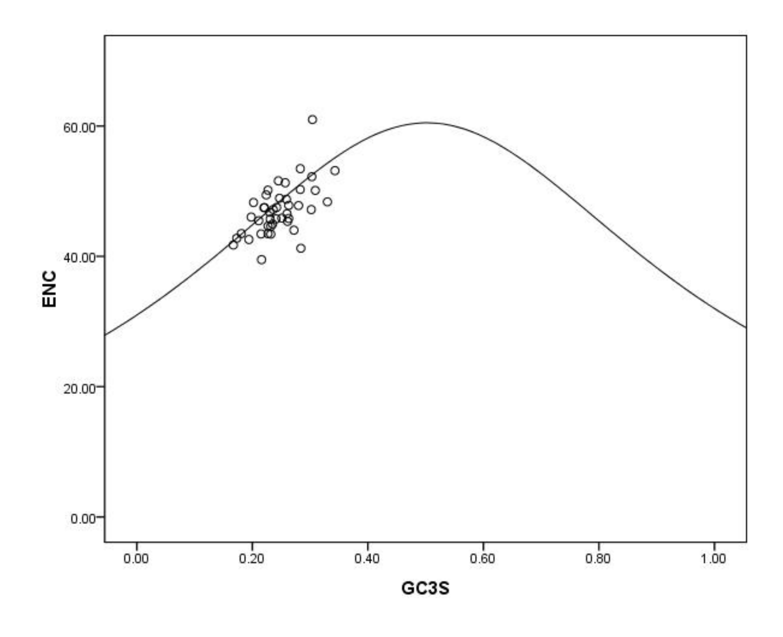


图 2 巨桉叶绿体基因的 ENC-plot 分析

Fig.2 ENC-plot analysis for genes on chloroplast genome in *Eucalyptus grandis*

表 4 巨桉叶绿体基因 ENC 比值频数分布

Table 4 ENC ratio of genes of chloroplast genome in *Eucalyptus grandis*

分组 Group	数目 Number	频率/% Frequency
-0.25~ -0.15	1	2.33
-0.15~ -0.05	4	9.30
-0.05~ 0.05	22	51.16
0.05~ 0.15	15	34.88
0.15~ 0.25	1	2.33

2.4 对应分析

对应分析表明，第 1 轴贡献率为 17.68%，第 2 轴贡献率为 11.49%，第 3、4 轴的贡献率分别 8.00%和 5.76%。前 4 个向量的总贡献率为 42.93%。第 1 轴和第 2 轴的贡献率均超过 10%，说明第 1 轴和第 2 轴都是密码子偏好的主要影响因素。第 1 轴与与 GC、 CAI、 CBI 和 Fop 呈极显著的正相关，其相关系数分别为 0.573、0.670、0.578 和 0.523；第 1 轴 ENC 呈极显著的负相关，其相关系数为-0.395；第 1 轴与 GC3S 无显著相关，而与第 3 位同义密码子 A 和 G 含量呈极显著相关，其相关系数分别为-0.440 和-0.606。为了更加直观的观察密码子偏好，建立以第 1 轴为 *x* 轴，以第 2 轴为 *y* 轴的平面坐标系，将所有参试基因按不同功能分布于坐标系中(图 3)。图 3 显示核糖体蛋白基因分布相对集中，其余基因分布相对比较分散，说明核糖体蛋白基因的密码子偏好相近，与其他基因的密码子偏好相差较大。

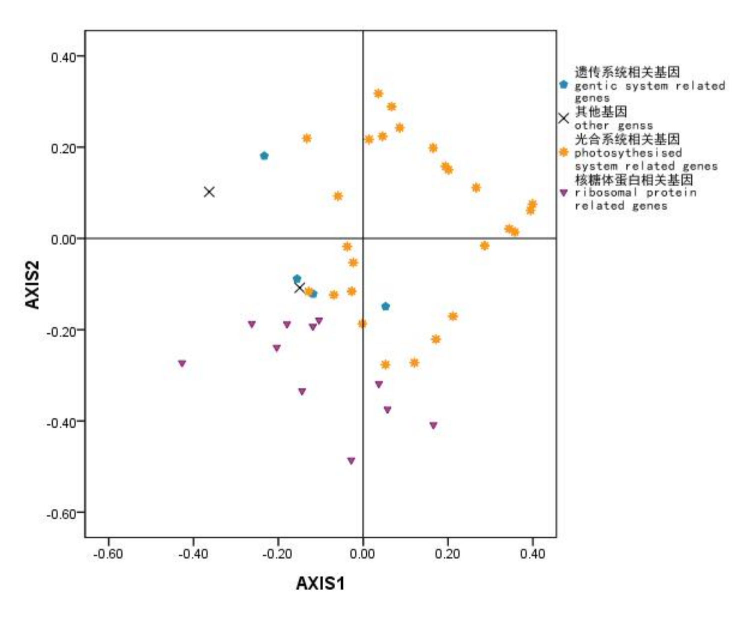


图 3 基于 RSCU 的对应性分析

Fig.3 Corresponding analysis of RSCU

2.5 最优密码子的确定

本文以密码子的 ENC 参数为标准，对参试基因进行排序，从两端各选取 10%的基因（本文中两端各选取 4 个），分别建立高/低表达基因库；在此基础上重新计算各表达库的 RSCU，求算出两个库的ΔRSCU（表 5）。以ΔRSCU>0.08 为标准确定了 31 个高表达密码子（表 5 中*标注的密码子），其中 12 以 G 结尾，8 个以 C 结尾，6 个以 A 结尾，5 个以 U 结尾。

将表 3 中的高频密码子与表 5 中确定的高表达密码进行分析，选取其中共有的密码子作为最优密码子。巨桉叶绿体基因中有 12 个最优密码子为 UUG、CUU、GUU、UCC、UCA、

ACA、UAU、UAA、CAU、AAU、AGA 和 GGA，其中 10 个密码子以 U 或 A 结尾，另外 2 个以 G 或 C 结尾。

表 5 巨桉叶绿体基因高/低表达库的同义密码子相对使用度

Table 5 Relative Synonymous Codon Usage of genes of chloroplast genome in *Eucalyptus grandis*

氨基酸 Amino acid	密码子 Codon	高表达基因 High expressed genes 数目 Number	高表达基因 High expressed genes RSCU	低表达基因 Low expressed genes 数目 Number	低表达基因 Low expressed genes RSCU	ΔRSCU
Phe	UUU	102	1.08	27	1.04	0.04
	UUC	87	0.92	25	0.96	-0.04
Leu	UUA	51	1.02	23	1.97	-0.95
	UUG*	69	1.38	13	1.11	0.27
	CUU**	80	1.61	15	1.29	0.32
	CUC**	27	0.54	2	0.17	0.37
	CUA	43	0.86	14	1.2	-0.34
	CUG**	29	0.58	3	0.26	0.32
Ile	AUU	103	1.33	41	1.64	-0.31
	AUC*	55	0.71	13	0.52	0.19
	AUA*	75	0.97	21	0.84	0.13
Met	AUG	62	1	34	1	0
Val	GUU*	51	1.5	22	1.35	0.15
	GUC*	20	0.59	6	0.37	0.22
	GUA	40	1.18	31	1.91	-0.73
	GUG**	25	0.74	6	0.37	0.37
Ser	UCU	77	1.64	18	1.89	-0.25
	UCC*	52	1.11	8	0.84	0.27
	UCA***	55	1.17	5	0.53	0.64
	UCG	39	0.83	8	0.84	-0.01
Pro	CCU	41	1.3	27	2.63	-1.33
	CCC***	32	1.02	2	0.2	0.82
	CCA	34	1.08	11	1.07	0.01
	CCG**	19	0.6	1	0.1	0.5
Thr	ACU	35	1.13	27	2.04	-0.91
	ACC	25	0.81	10	0.75	0.06
	ACA**	43	1.39	13	0.98	0.41
	ACG**	21	0.68	3	0.23	0.45
Ala	GCU	42	1.75	39	2.44	-0.69
	GCC*	13	0.54	4	0.25	0.29
	GCA	25	1.04	17	1.06	-0.02
	GCG**	16	0.67	4	0.25	0.42
Tyr	UAU*	88	1.59	24	1.33	0.26
	UAC	23	0.41	12	0.67	-0.26
TER	UAA***	3	2.25	2	1.5	0.75
	UAG	1	0.75	2	1.5	-0.75

His	CAU**	54	1.52	10	1.05	0.47
	CAC	17	0.48	9	0.95	-0.47
Gln	CAA	87	1.4	19	1.73	-0.33
	CAG**	37	0.6	3	0.27	0.33
Asn	AAU**	132	1.52	21	1.17	0.35
	AAC	42	0.48	15	0.83	-0.35
Lys	AAA	114	1.25	19	1.73	-0.48
	AAG**	69	0.75	3	0.27	0.48
Asp	GAU	135	1.68	15	1.67	0.01
	GAC	26	0.32	3	0.33	-0.01
Glu	GAA	116	1.25	35	1.56	-0.31
	GAG**	69	0.75	10	0.44	0.31
Cys	UGU	22	1.29	13	1.86	-0.57
	UGC***	12	0.71	1	0.14	0.57
TER	UGA	0	0	0	0	0
Trp	UGG	57	1	20	1	0
Arg	CGU	20	0.68	22	2.4	-1.72
	CGC	11	0.37	8	0.87	-0.5
	CGA	44	1.49	13	1.42	0.07
	CGG***	18	0.61	0	0	0.61
Ser	AGU	43	0.91	14	1.47	-0.56
	AGC	16	0.34	4	0.42	-0.08
Arg	AGA***	53	1.8	10	1.09	0.71
	AGG***	31	1.05	2	0.22	0.83
Gly	GGU	33	0.95	43	2.39	-1.44
	GGC*	17	0.49	7	0.39	0.1
	GGA***	61	1.76	18	1	0.76
	GGG***	28	0.81	4	0.22	0.59

3.讨论

遗传密码是指核苷酸序列与氨基酸序列的对应关系。20 种蛋白质氨基酸中 Met 和 Trp 两种氨基酸只有一个密码子，其余 18 种氨基酸均有 2~6 个不等密码子编码，即密码子的简并性，编码同一氨基酸的密码子为同义密码子(朱圣庚和徐长发, 2016) 。同义密码子差别主要在于第三位密码子的变化。本文中巨桉叶绿体基因组中 GC3 与 GC1 和 GC2 无显著相关，并且明显小于 GC1 和 GC2。这说明巨桉叶绿体基因密码子偏好以 A 和 U 结尾，RSCU 分析结果从定量分析的角度也充分证明这一观点。这与已报道的黄芩(*Scutellaria baicalensis*)(王文斌等, 2018)，普通油茶(*Camellia oleifera*)(王鹏良等, 2018)，蒺藜苜蓿(*Medicago truncatula*)(杨国锋等, 2015)等植物叶绿体基因的特征一致。

生物在编码氨基酸时经常倾向使用某个特定的同义密码子的现象称为密码子使用偏好性(吴宪明等, 2007)。巨尾桉叶绿体基因组的密码子的 ENC 平均值为 47.04。以 35 为标准，ENC 低于 35 的为强偏好性密码子；高于 35 的为弱偏好性密码子(Jiang, 2008)。因此，巨尾桉叶绿体基因组密码子为弱偏好性的，CAI 参数也支持这一观点。

密码子偏好受碱基组成、选择、tRNA 丰度、基因长度和蛋白质的疏水性等许多因素的影响 (梁菲菲, 2010)。本文中由于密码子数目不与 ENC 及 CAI 等密码子参数显著相关，所

以巨桉叶绿体基因组中基因长度对密码子偏好没有明显作用；同样蛋白质疏水性对巨桉叶绿体基因组密码子偏好也无明显作用；同义密码子第 3 位 A 和 G 含量及 GC 含量与第 1 轴达到极显著相关，表明碱基差异对密码子的偏好也有影响；CAI、CBI 和 FOP 与第 1 轴极显著相关，表明基因有选择地使用高丰度的 tRNA 对应的密码子，导致基因的高表达(Ikemura, 1981a; 1981b;1985)，所以选择也是影响密码子偏好的重要原因；由于第 3 位 A、G 含量和 GC 含量与第 1 轴的相关系数与 CAI、CBI 和 FOP 和第 1 轴的相关系数比较接近，因此突变和自然选择在密码子偏好中的作用基本相当。中性绘图分析表明选择是导致密码子偏好的相对主要因素；而 ENC-plot 分析结果显示突变也占较大比例。因此，本文认为突变和选择在巨桉叶绿体基因组中可能起相对均衡的作用。

本文以高表达的高频密码子为最优密码子，在巨桉叶绿体基因组中确定的 12 个最优密码子分别为 UUG、CUU、GUU、UCC、UCA、ACA、UAU、UAA、CAU、AAU、AGA 和 GGA。巨桉叶绿体基因组最优密码子的确定为优化目标基因的密码子，提高表达效率，从而利用叶绿体基因工程改良巨桉重要性状奠定良好基础。

参考文献

CHEN SX, ZHENG JQ, LIU XF,et al, 2018. Hundred year histories and prospect of Eucalyptus cultivation technology development in China[J]. World Forestry Research, 31:7-21. [陈少雄, 郑嘉琪,刘学锋, 2018. 中国桉树培育技术百年发展史与展望[J]. 世界林业研究，31:7-12.]

DANIELL H, CHASE C,2004. Molecular biology and biotechnology of plant organelles[M]. Dordrecht: Springer.

HANSON G, COLLER J, 2018. Codon optimality, bias and usage in translation and mRNA decay[J]. Nat Rev Mol Cell Biol, 19:20-30.

HERSHBERG R, PETROV DA, 2008. Selection on codon bias[J]. Annu Rev Genet, 42:287-299.

IKEMURA T, 1981a. Correction between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the Respective codons in its protein genes[J]. J Mol Biol, 146:1-21.

IKEMURA T, 1981b. Correction between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codon in its protein genes : A proposal for a synonymous codon choice that is optimal for the *E. coli* translation system[J]. J Mol Biol, 151:389-409.

IKEMURA T, 1985. Codon usage and tRNA content in unicellular and multicellular organisms[J]. Mol Biol Evol 2:13-34.

JIANG Y, DENG F, WANG H, et al., 2008. An extensive analysis on the global codon usage pattern of baculoviruses[J]. Arch Virol, 153:2273-2282.

LI J, XUE QZ, 2005. Comparison of MADS transcriptional factor on codon bias in arabidopsis and rice[J]. J Zhejiang Univ (Agric Life Sci Ed): 513-517. [李娟,薛庆中, 2005. 拟南芥及水稻转录因子 MADS 密码子的偏好性比较[J]. 浙江大学学报(农业与生命科学版):513-517.]

LIANG FF, 2010. Influencing of codon bias and its research significance [J]. Anim Husb Feed Sci, 31(1):118-119. [梁菲菲，2010.密码子偏性的影响因素及研究意义[J].畜牧与饲料科学 31(1):118-119.]

LIU H, WANG MX, YUE WJ, 2017. Analysis of codon usage in the chloroplast genome of Broomcorn millet (*Panicum miliaceum* L.) [J]. Plant Sci J, 35:362-371 [刘慧，王梦醒，岳文杰，等, 2017. 糜子叶绿体基因组密码子使用偏性的分析[J]. 植物科学学报 35:362-371.]

LIU J, XIANG DY, CHEN JB, et al., 2009. Low temperature LT50 of three Eucalyptus seedlings with Electrical conductivity method and logistic equation[J]. Guangxi Forestry Science, 38:75-78. [刘建，项东云，陈健波，等, 2009. 应用 Logistic 方程确定三种桉树的低温半致死温度[J]. 广西林业科学 38:75-78.]

NELSON DL, COX MM, 2017. Lehninger Principles of Biochemistr[M]. New York: W.H.Freeman and Company

QI SX, 2006. Introduction and status of Eucalyptus in China[J].Guangxi Forestry Science, 35:250-252. [祁述雄, 2006. 中国引种桉树与发展现状[J]. 广西林业科学,35:250-252]

WANG GL, FANG HJ, 2014. Plant genetic engineering[M]. Beijing: Science Press [王关林,方宏筠,2014. 植物基因工程[M]. 北京：科学出版社.]

WANG PL, YANG LP, WU HY, et al, 2018. Codon preference of chloroplast genome in *Camellia oleifera*[J]. Guihaia, 38:135-144 [王鹏良，杨利平，吴红英，等, 2018. 普通油茶叶绿体基因组密码子偏好性分析[J]. 广西植物，38:135-144.]

WANG WB, YU H, QIU XP, 2018. Analysis of repeat sequence and codon bias of chloroplast genome in *Scutellaria baicalensis*[J]. Molecular Plant Breeding, 16:2445-2452. [王文斌，于欢,邱相坡, 2018. 黄芩叶绿体基因组重复序列及密码子偏好性分析[J]. 分子植物育种 16:2445-2452.]

WRIGHT F, 1990. The effective number of codons used in a gene[J]. Gene, 87:23-29.

WU SJ, CHEN GC, XU JM, et al., 2016. Variation analysis and selection for *Eucalyptus grandis* provenances and families in multiple-sties[J]. For Environ Sci, 32: 10-15 [吴世军，陈广超，徐建民，等, 2016. 巨桉种源/家系多点遗传变异及选择比较[J]. 林业与环境科学，32:10-15]

WU XM, WU SF, REN DM, et al, 2007. The analysis method and progress in the study of codon bias[J].HEREDITAS, 29:420-426. [吴宪明，吴松锋，任大明等, 2007. 密码子偏性的分析方法及相关研究进展[J]. 遗传, 29:420-426.]

XU C, BEN AL, CAI XN, 2010. Analysis of synonymous codon usage in chloroplast geneome of *Phalaenopsis aphrodite* subsp. *formosana*[J]. Mol Plant Breed, 8:945-950 [续晨，贲爱玲，蔡晓宁, 2010. 蝴蝶兰叶绿体基因组密码子使用的相关分析[J]. 分子植物育种 8:945-950.]

YANG GF, SU KL, ZHAO YR, et al., 2015. Analysis of codon usage in the chloroplast genome of *Medicago truncatula*[J].Acta Prataculturae Sinica, 24:171-179. [杨国锋，苏昆龙，赵怡然,等, 2015. 蒺藜苜蓿叶绿体密码子偏好性分析[J]. 草业学报 24:171-179.]

ZHANG J, CHEN GC, XU JM, et al., 2016. Comprehensive selection for *Eucalyptus grandis* provenances and families[J]. Journal of Tropical and subtropical Botany, 24:280-286. [张捷，陈广超，徐建民，等, 2016. 巨桉种源/家系综合选择研究[J]. 热带亚热带植物学报，24:280-286.]

ZHANG ZY, XIANG DY, XU JM, et al., 2016. Comprehensive analysis of growth, stem form and

resistance to *Leptocybe invasa* of *Eucalyptus grandis* provenances[J]. Forest Resources Management:107-111 [张照远，项东云，徐建民,等, 2016. 不同种源巨桉生长、干形和抗桉树枝瘿姬小蜂的综合评价[J]. 林业资源管理:107-111.]

ZHOU M, LONG W,LI X, 2008. Analysis of synonymous codon usage in chloroplast genome of *Populus alba*[J]. J For Res, 19:293-297.

ZHOU ZP, DANG YK, ZHOU M, et al, 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription[J]. Proc Nat Acad Sci USA 26:e6117-e6125.

ZHU SG, XU CF, 2016. Biochemistry (4th Ed) [M]. Beijing: Higher Education Press. [朱圣庚,徐长发,2016. 生物化学(第四版)[M]. 北京：高等教育出版社.]